INFORMATION AND CONTROL 7, 416–444 (1964)

Nonsupervised Adaptive Signal Detection and Pattern Recognition

DAVID B. COOPER

Columbia University, New York, New York
and
Raytheon Company, Norwood, Massachusetts

AND

PAUL W. COOPER

Applied Research Laboratory, Sylvania Electronic Systems, Waltham, Massachusetts

Adaptive signal detection and pattern recognition can be viewed as a problem in statistical classification wherein the partitioning of an n-dimensional sample space into category (signal) regions is determined through estimation from a set of samples from the categories. When the correct associations of the samples are known, the problem is the commonly treated *supervised* one. This paper, examining the *nonsupervised* case wherein the correct associations of the samples are unknown, demonstrates that it is possible under extremely general conditions to achieve effective adaptation without supervision. With particular emphasis on a two-category (binary detection) model, general conditions are described under which nonsupervised adaptation is possible, and specific simple yet rapidly convergent techniques are presented under varying degrees of prior knowledge of the statistical properties of the data.

Most of the paper is concerned with a two-category case where the corresponding (equiprobable) distributions differ only in location. The paper proceeds by examining the over-all probability distribution comprised of the two component category distributions, and the adaptation treated is directed toward determining the decision boundary, or the distribution parameters necessary for defining it. For univariate normal distributions various estimators (and their convergence properties) of the over-all mean are examined. For multivariate *monotone* (including normal) distributions the over-all sample covariance matrix is used to obtain the component covariance matrices when these are general (including the colored noise case), or simply to obtain the principal eigenvector (of the over-

all matrix) when the component distributions are spherically symmetric (white noise). A hill-climbing algorithm is included. These results for the important model of binary signal detection in gaussian noise demonstrate that no prior knowledge of the signal or noise parameters is required for nonsupervised adaptation to the optimum detector. It is shown in one dimension that for equiprobable component distributions of almost any functional form and differing only by translation we can obtain category distribution estimators which converge uniformly over the real line with probability 1. Considered also are the case of different a priori probabilities, the problem of tracking, and some aspects of the multiple-category problem.

I. INTRODUCTION

Pattern recognition can be viewed as a problem in statistical classification wherein an n-dimensional sample space is partitioned into category regions with decision boundaries. Members of the categories are each represented by a sequence of n numbers, or equivalently as a vector in the n-dimensional hyperspace. Assuming there exists a probability distribution associated with each category describing the distribution of its members in n-space, the object is to partition the space in an optimal fashion. An unknown vector is then assigned to the category in whose region it falls. This is the classic model on which so many of the statistical pattern recognition studies have been based.

Signal detection can be treated in terms of this pattern recognition model, a time waveform being represented by a discrete set of samples according to any one of the many sampling theorems (Shannon, 1949). Included here are cases where deterministic signals are corrupted by additive noise, cases where the signals themselves are stochastic, and more generally where the channel too introduces nonlinear stochastic transformations. But examination of the received waveforms reduces to the general classification problem already described. In the general pattern recognition problem the distributions exist because of inherent differences among the members of each of the various categories, whereas in the signal detection case the channel itself introduces some of these differences. Examination of channel properties (not part of this paper) serves primarily to introduce further knowledge of the statistics of the received samples.

The concepts of adaptation have been developed under certain restrictive conditions. On the basis of a sequence of samples from the categories (signals) and any a priori knowledge of the probability structure of the

problem, the partitions (decision rules) are estimated. But this work has been generally predicated upon the assumption that the correct associations of the "learning" samples with their appropriate categories are known. That is, the adaptation is *supervised* in that there is a teacher directing the "learning." Eventually, in the recognition phase, the assignments are made on the basis of the already obtained partitions.

There are, however, situations for which supervision is impossible or inconvenient. The prospect of performing *nonsupervised* adaptation has recently fired the imagination of researchers. Here the samples on which adaptation is based are not associated a priori in any known way with the categories. Many people have felt this problem to be formidable and approachable only under somewhat limited conditions, perhaps following an initial period of supervised "learning." An approach to the problem was recently made by Jakowatz, Shuey, and White (1961), whose work was further examined by Hinich (1962). Under restrictive conditions the supervised requirement was relaxed in the earlier Rake system (Price and Green, 1958).

The present paper examines the problem of *nonsupervised* adaptation and demonstrates that it can be effectively achieved under a wide range of situations. Except where otherwise indicated it is assumed that the categories have equal a priori probabilities. The criterion we use to define *optimality* is that the total probability of misclassification be minimal. This is a Maximum-likelihood criterion corresponding to the Bayes criterion for equal costs of misclassification, and represents the Ideal Observer. However, when we actually estimate the distributions themselves, we can use any decision criterion—Bayes, Minimax, Neyman-Pearson, etc.—although the resultant decision boundary may then have a much altered form. The minimal-error-probability criterion is used throughout the paper assuming at any time that the parameter values are true. *Through convergent estimators we adapt to the decision boundary optimum for the true distributions, for which the irreducible error is achieved.*

Throughout the paper the words detector, decision rule, and decision boundary are used essentially interchangeably. This work as well as all the other statistical pattern recognition and signal detection work is of course built upon the foundations of multivariate discriminant analysis widely described in the mathematical statistical literature, e.g., (Wilks, 1962).

II. VIEWPOINT

If it is known, for example, that the categories are tightly clustered and widely separated, then simple approximative partitioning procedures become apparent. We have been interested in determining nonsupervised adaptive partitions convergent to the true optimal one, regardless of the degree of overlap of the category distributions.

Our viewpoint has been one in which the individual category distributions comprise the "modes" of an over-all "multimodal" distribution, from which the samples are considered to have been drawn. Attention is directed at estimating, explicitly or implicitly, parameters of this "multimodal" distribution defining the partition. As is to be expected, nonsupervised adaptation cannot be uniquely achieved for arbitrary distributions. But where there is adequate probability structure to the problem, the partition can be unique. There are many cases for which this is possible. Of special importance is the two-category case where the distributions are translates of one another, but have general functional form, and further interest centers on the cases where the distributions are finitely parameterized. Much simplification is achieved by only partial prior knowledge of the probability structure of the problem, e.g., the mean of one of the distributions. However, major interest in this paper centers on the situations where there is no prior knowledge of any of the distributional parameters.

III. ONE-DIMENSIONAL NORMAL DISTRIBUTIONS

In this section we treat the problem where we have m samples each from one of two categories having (initially unknown) univariate normal distributions differing only in their means. From these samples we wish to determine the decision threshold, which for the minimum-error-probability criterion is the mean of the two means. The two population distributions are:

$$p_j(x) = N(u_j, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(1/2\sigma^2)(x-u_j)^2}, \qquad j = 1, 2. \quad (1)$$

In the usual *supervised* approach it is known that particular sets of m_j samples $\{z_k^{(j)}\}$ are from the jth category, where $m = m_1 + m_2$. The category means are estimated with the sample mean,

$$\hat{u}_j = \frac{1}{m_j} \sum_{k=1}^{m_j} z_k^{(j)}, \quad (2)$$

and the threshold is estimated

$$\hat{\mu} = \tfrac{1}{2}(\hat{u}_1 + \hat{u}_2). \tag{3}$$

The estimator $\hat{\mu}$ is unbiased with variance

$$\text{Var} = \frac{\sigma^2}{m}\left[\frac{m^2}{4m_1(m - m_1)}\right]. \tag{4}$$

Considering that the a priori distribution of m_1 is binomial, the expected variance of $\hat{\mu}$ is obtained by averaging (4) according to m_1 so distributed (where for $m_1 = 0$ or $m_1 = m$ we average in the value $\sigma^2 + \alpha^2$, this contribution vanishing for large m). The expected variance is greater than σ^2/m, reaching it asymptotically as m gets large.

NONSUPERVISED

The purpose of this paper is to show how effectively the problem can be tackled when we know nothing of the identity of the samples,but merely that we have m samples drawn in some fashion from the two categories. We can view the problem as one in which we have m samples from a bimodal distribution for which we wish to estimate the mean μ, which is what in effect was done in the supervised case. Expressed in terms of its component (unimodal) normal distributions, the resultant bimodal distribution is

$$\begin{aligned} p(x) &= \tfrac{1}{2}[N(u_1, \sigma^2) + N(u_2, \sigma^2)] \\ &= \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(1/2\sigma^2)\alpha^2} e^{-(1/2\sigma^2)(x-\mu)^2} \cosh[(\alpha/\sigma^2)(x - \mu)], \end{aligned} \tag{5}$$

where $\mu = u_2 - \alpha = u_1 + \alpha$. The term bimodal is loosely used here to describe an over-all distribution comprising two distinct unimodal distributions. Strictly speaking, the bimodal form of the over-all distribution does not become evident unless $|u_2 - u_1| > 2\sigma$, for component normal distributions.

In terms of the m samples $\{z_k\}$, the simplest estimate of the mean is the sample mean

$$\hat{\mu} = \frac{1}{m}\sum_{k=1}^{m} z_k . \tag{6}$$

The distribution of this unbiased estimator is asymptotically normal, and for any m its variance is

$$\text{Var} = (1/m)(\sigma^2 + \alpha^2) = (\sigma^2/m)(1 + \gamma^2), \tag{7}$$

where $\gamma = |\alpha|/\sigma$ is a signal-to-noise ratio, which in effect is a measure of the separation of the category means in units of standard deviation. Measures of performance (probabilities of error) are all functions specifically of γ.

The sample mean (6) converges for *all* values of signal-to-noise ratio, and it is especially effective for small values. In fact, for $\gamma < 1$ the sample mean has variance little different from the minimum variance achieved asymptotically by the maximum-likelihood estimator. For $\gamma < \frac{3}{4}$, this difference becomes negligible.

It is interesting, and at first surprising, to note that for finite m and small γ the (nonsupervised) sample mean can be better than the usual supervised solution of Eq. (3). For example, as γ goes to zero, the overall distribution becomes in effect a single unimodal normal, for which we desire the mean. The best estimator is then the sample mean. For given m_1 and m_2 not equal, the usual supervised solution has a greater variance, as given in (4). Its expected variance is also greater than that of the sample mean, only reaching it asymptotically as m goes to infinity. The variance of the sample mean can be less than the supervised also for γ greater than zero, but small; although for $\gamma > 0$ the limiting expected supervised variance will be minimal.

The sample mean lends itself simply for adaptive updating wherein all that need be remembered is the latest estimate and the number m. Denoting by $\hat{\mu}_m$ the estimate based on m samples, an additional sample z_{m+1} is incorporated,

$$\hat{\mu}_{m+1} = \frac{m\hat{\mu}_m + z_{m+1}}{m+1}. \tag{8}$$

Although the maximum-likelihood estimator is not readily obtained in explicit form, its examination is of interest. For all parameters unknown, the solutions are discussed in Appendix I. For only μ unknown, the maximum-likelihood estimator is obtained from solution of (9).

$$\hat{\mu} = \frac{1}{m}\sum_{k=1}^{m} z_k - \frac{\alpha}{m}\sum_{k=1}^{m} \tanh\left[(\alpha/\sigma^2)\,(z_k - \hat{\mu})\right]. \tag{9}$$

The asymptotic variance for the maximum-likelihood estimator is derived in Appendix II, and is portrayed graphically in terms of γ in Fig. 1. One observes that for small γ or for very large γ the asymptotic variance approaches σ^2/m, which is as good as is obtained asymptotically for the fully supervised case. And even for the least favorable γ, the variance is less than $2\sigma^2/m$.

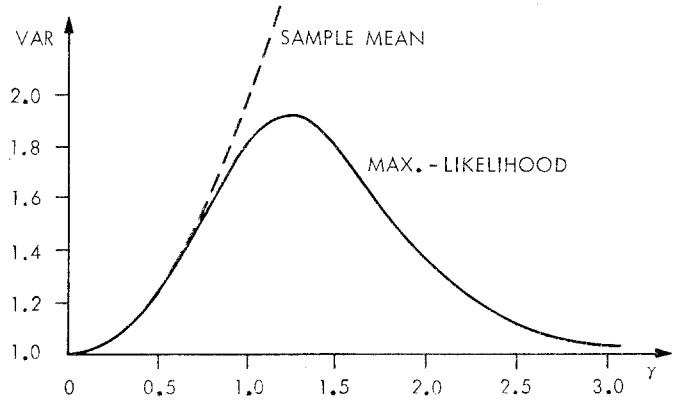

FIG. 1. Variance curves (versus γ) in units of σ^2/m

For small γ the sample mean is near optimum. Although it is also good for large γ, we could do better there with the maximum-likelihood estimator. Although the latter cannot be conveniently obtained explicitly, for large γ it can be closely approximated rather easily. For large γ,

$$\tanh [(\alpha/\sigma^2)(z_k - \hat{\mu})] \text{ is } \pm 1 \text{ with high probability.}$$

Equation (9) is then effectively

$$\hat{\mu} = \frac{1}{m} \sum_{k=1}^{m} z_k - \frac{|\alpha|}{m} H, \tag{10}$$

where H is the difference between the number of $\{z_k\}$ greater than $\hat{\mu}$ and the number less than $\hat{\mu}$. H can be determined by remembering all $\{z_k\}$ and by comparing them with

$$\hat{\mu}' = \frac{1}{m} \sum_{k=1}^{m} z_k .$$

When α is not known, then it can be estimated from

$$|\hat{\alpha}| = \frac{1}{m-1} \sum_{k=1}^{m} |z_k - \hat{\mu}'| ,$$

which is then used in (10).

A convenient measure for deciding whether the signal-to-noise ratio is large is the kurtosis β, expressed simply in terms of the second and fourth central moments.

$$\beta = \frac{E[(x-\mu)^4]}{\{E[(x-\mu)^2]\}^2} . \tag{11}$$

For the normal component distributions treated here,

$$\beta = 1 + 2(1 + 2\gamma^2)(1 + \gamma^2)^{-2} = 3 - 2\gamma^4(1 + \gamma^2)^{-2} \quad (12)$$

Inversion of (12) leads to

$$\gamma^2 = \frac{3 - \beta}{\beta - 1}\left[1 + \sqrt{2/(3 - \beta)}\right]. \quad (13)$$

Figure 2 portrays β versus γ. Strictly speaking the kurtosis defined in (11) is used to indicate an aspect of the shape of a unimodal distribution, and our application of it here to a bimodal distribution is one of mathematical convenience, since it is a measure of γ, dependent only upon the relative values of α and σ, not on their specific values. The first, second, and fourth sample moments are used for estimating β, and the simplicity inherent in updating of sample moments, as in (8) for the sample mean, is achieved.

The estimator of (10) introduces a correction term to the sample mean and thereby tends to remove the uncertainty arising from the disparity in the division of the samples between the categories. A variation on (10) is to determine the sample mean; then having retained all of the samples, obtain the mean of all of the samples on each side, and take as revised threshold the mean of these two means, thereby simulating a supervised solution. This latter procedure can be effected sequentially whereby each sample as it occurs is assigned and then used to modify the threshold. In one form or another such a nonlinear approach has been taken by a number of researchers. At any given stage of adaptation the threshold value for the sequential nonlinear method

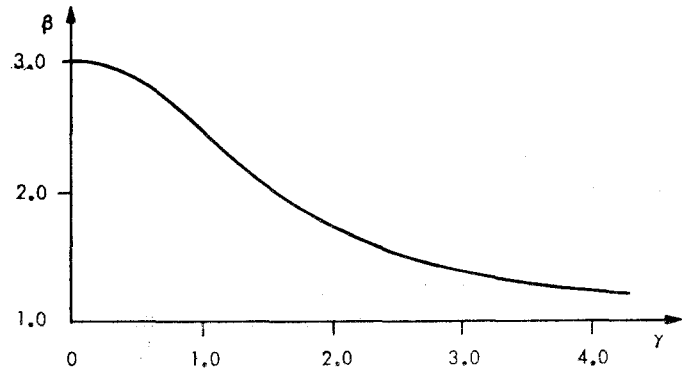

FIG. 2. β versus γ for normal distributions

is a function of the order in which the samples occurred (or were processed) in contrast to the linear estimation (of the updated sample mean) for which such ordering is irrelevant. The nonlinear sequential method could proceed in several possible ways for initially determining the threshold. Initially the threshold could be the mean of the first two samples. Or it could be taken as the mean of the first K samples (or perhaps of their two extreme members). Thereafter the two running means determine the updated threshold. Transients arising because of incorrect assignment of early samples can be discounted by weighted averaging.

The variance of the threshold estimator is of course important only to the extent it degrades the performance, in that the uncertainty in the threshold increases the expected probability of error. For large γ the (discrimination) error probability is negligible, and the larger variance of some threshold estimators may not be important. Comparison of the variances of the various estimators then serves merely to order their relative values, although perhaps the differences in actual performance may be marginal. But if the best performance is required, then what is of interest is the actual degradation caused by the threshold uncertainty, and this can be determined.

Before concluding our discussion of the univariate case, we comment on the sample median. The well known relation for the asymptotic variance of the sample median is $[4mp^2(\mu)]^{-1}$. Substitution of (5) leads to determination of the ratio of the asymptotic variance of the sample median to the variance of the sample mean. This relative variance is

$$R = \frac{\pi}{2} \frac{e^{\gamma^2}}{(1 + \gamma^2)}. \tag{14}$$

At $\gamma = 0$ it is $\pi/2$, as expected for the normal distribution. It increases to infinity with γ, but the sample median is nevertheless convergent for all γ.

IV. n-DIMENSIONAL NORMAL DISTRIBUTIONS

Two spherically symmetric multivariate normal distributions differing only in location are of course optimally partitioned with a hyperplane which is the perpendicular bisector of the line connecting the two means. Defining this plane by a vector $\mathbf{v}$ perpendicular to it and by a point in it, e.g., $\mathbf{u}$, an unknown $\mathbf{x}$ is assigned to one of the categories according as $(\mathbf{v}'\mathbf{x} - \rho)$ is greater or less than zero, where the threshold ρ

is equal to $\mathbf{v}'\boldsymbol{\mu}$. Such a decision procedure could of course be achieved by passing signal $\mathbf{x}$ through a filter having reverse impulse response $\mathbf{v}$, and comparing the output with threshold ρ, where time function equivalents of these vectors could be considered. Although $\mathbf{v}$ can be normalized, it can be defined in terms of the component means as $\mathbf{v} = (\mathbf{u}_2 - \mathbf{u}_1)$, and $\rho = (\frac{1}{2})(\mathbf{u}_2'\mathbf{u}_2 - \mathbf{u}_1'\mathbf{u}_1)$. A *supervised* solution is obtained from the sample means of the two component distributions.

As in the univariate case, the *nonsupervised* approach involves examination of the over-all bi-modal distribution, which for component normal distributions is

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\,\sigma^n}\{\exp[-(1/2\sigma^2)\boldsymbol{\alpha}'\boldsymbol{\alpha}]\}\cdot\{\exp[-(1/2\sigma^2)(\mathbf{x}-\boldsymbol{\mu})' \cdot(\mathbf{x}-\boldsymbol{\mu})]\}\cosh[(1/\sigma^2)\boldsymbol{\alpha}'(\mathbf{x}-\boldsymbol{\mu})] \tag{15}$$

where, analagous to the univariate case, $\boldsymbol{\mu} = \mathbf{u}_1 + \boldsymbol{\alpha} = \mathbf{u}_2 - \boldsymbol{\alpha}$. (Here, the indexing of $\mathbf{u}_1$ and $\mathbf{u}_2$ is arbitrary, since all that we know is that there are two component distributions. Therefore any relationships we develop would not prescribe the algebraic sign for $\boldsymbol{\alpha}$, i.e., they are applicable to $\pm\boldsymbol{\alpha}$. Therefore, for convenience, we shall arbitrarily select the first coordinate of $\boldsymbol{\alpha}$, that is α_1, as being positive, and the relationships of the other coordinates follow.) (We might note that in practice in dealing with estimated parameters, it would be better to instead arbitrarily treat the largest α_i as being positive, and this in fact should be assumed wherever α_1 is used as a reference for determining the other coordinates of $\boldsymbol{\alpha}$.) *What the problem boils down to then is estimation of the mean* $\boldsymbol{\mu}$ *and determination of the principal axis of the bimodal distribution. The latter is equivalent to determination of the principal component or to determination of the eigenvector corresponding to the largest eigenvalue of the over-all covariance matrix.*

The sample mean is determined

$$\hat{\boldsymbol{\mu}} = \frac{1}{m}\sum_{k=1}^{m}\mathbf{z}_k. \tag{16}$$

An especially simple situation arises when the mean of one of the component distributions is known, say $\mathbf{u}_1$. Then $\hat{\mathbf{v}}$ is simply $(\hat{\boldsymbol{\mu}} - \mathbf{u}_1)$. Of course $\hat{\mathbf{u}}_2$ is then $(2\hat{\boldsymbol{\mu}} - \mathbf{u}_1)$. When $\mathbf{u}_1 = 0$, this reduces to the signal detection problem wherein we wish to decide whether or not signal is present.

For the general case where both category means are unknown (as is σ), we estimate the over-all covariance matrix $\mathbf{B}$, where

$$\hat{\mathbf{B}} = \frac{1}{m-1} \sum_{k=1}^{m} (\mathbf{z}_k - \hat{\boldsymbol{\mu}})(\mathbf{z}_k - \hat{\boldsymbol{\mu}})'. \tag{17}$$

If the component distributions each had a general covariance matrix $\boldsymbol{\Sigma}$, then in terms of the vector $\boldsymbol{\alpha}$,

$$\mathbf{B} = \boldsymbol{\alpha}\boldsymbol{\alpha}' + \boldsymbol{\Sigma}, \tag{18}$$

or in terms of the matrix elements and the vector components,

$$b_{ij} = \alpha_i \alpha_j + \sigma_{ij}. \tag{19}$$

For the spherically symmetric case treated in this section, $\boldsymbol{\Sigma}$ is diagonal with equal elements σ^2. The discussion here is in terms of the relationships among the true parameters, and, in practice, parameters determined in terms of the sample mean and sample covariance matrix should of course be overscored with circumflex to indicate that they are estimates.

One would like to avoid the formidable task of solving the characteristic equation for all of the eigenvalues, selecting the largest, and determining its corresponding eigenvector by solving n simultaneous equations. Classic iterative techniques (Faddeeva, 1959) lead directly to (generally rapidly convergent) simultaneous determination of the largest eigenvalue and its eigenvector. However, we can dispense with the above and instead take advantage of the special symmetry of the ellipsoid representing the matrix $\mathbf{B}$, for which there is a largest eigenvalue λ_1, and where the remaining ones are all equal and of value λ_2. We then note that

$$\begin{aligned} \lambda_1 &= \boldsymbol{\alpha}'\boldsymbol{\alpha} + \sigma^2 \\ \lambda_2 &= \sigma^2. \end{aligned} \tag{20}$$

To determine σ we obtain λ_2, and this can be expressed in terms of the coefficients of the characteristic equation which is an nth order polynomial equation

$$\sum_{k=0}^{n} (-1)^{n-k} T_k \lambda^k = 0. \tag{21}$$

The coefficients $\{T_k\}$ are invariant under diagonalization of $\mathbf{B}$, and they are each sums of determinants of some of the submatrices of $\mathbf{B}$; e.g.,

$T_n = 1$ and $T_0 = |\mathbf{B}|$. The ones of special interest to us are T_{n-1} and T_{n-2}. T_{n-1} is the Trace, whereby

$$T_{n-1} = \sum_{i=1}^{n} b_{ii} \tag{22}$$

and

$$T_{n-1} = \lambda_1 + (n-1)\lambda_2 . \tag{23}$$

T_{n-2} is the sum of all 2×2 determinants defined on the diagonal,

$$T_{n-2} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (b_{ii}b_{jj} - b_{ij}^2) = \tfrac{1}{2}\sum_{i=1}^{n} \sum_{j=1}^{n} (b_{ii}b_{jj} - b_{ij}^2), \tag{24}$$

and

$$T_{n-2} = (n-1)\lambda_1\lambda_2 + (\tfrac{1}{2})(n-1)(n-2)\lambda_2^2. \tag{25}$$

Solution of (23) and (25) leads to

$$\lambda_2 = \frac{T_{n-1}}{n}\left\{1 - \left[1 - \left(\frac{2n}{n-1}\right)\frac{T_{n-2}}{T_{n-1}^2}\right]^{1/2}\right\}. \tag{26}$$

From (19) and (20) we obtain $b_{ii} = \alpha_i^2 + \sigma^2 = \alpha_i^2 + \lambda_2$, or

$$\alpha_i^2 = b_{ii} - \lambda_2 , \qquad i = 1, 2, \cdots, n. \tag{27}$$

Taking α_1 positive, the algebraic signs of the $\{\alpha_i\}$ are obtained by examining the first column of $\mathbf{B}$ and choosing for α_i the same sign as b_{i1}. Then $\boldsymbol{\alpha}$ is determined, and $\mathbf{v}$ can be equated to $\boldsymbol{\alpha}$. Should we want to normalize $\mathbf{v}$ and obtain direction cosines, these are

$$l_i = \alpha_i \left(\sum_{j=1}^{n} \alpha_j^2\right)^{-1/2} . \tag{28}$$

There is an alternative and even simpler set of relationships for determining $\boldsymbol{\alpha}$, and we comment upon it here. For $n \geqq 3$, simultaneous solution of b_{21}, b_{31}, and b_{32} (where $b_{ij} = \alpha_i\alpha_j$) for α_1 gives

$$\alpha_1^2 = \frac{b_{21}\, b_{31}}{b_{32}} \tag{29}$$

where we take the positive root. Then α_i is

$$\alpha_i = b_{i1}/\alpha_1 . \tag{30}$$

(For $n = 2$, simultaneous solution of the three equations of (18) gives the desired parameters.) Various permutations of these relations lead

to other expressions. For example, each of the α_i can be found from expressions equivalent to (29), where (30) is then used only to determine sign. A disadvantage of this alternate method is that when we are working with estimated parameters rather than with the true ones, error in estimating α_1 is carried forth toward all the other estimates. And, in fact, if α_1 were not one of the larger components, the estimates of (30) could be greatly in error. Note that the main method finds λ_2, (26), in terms of coefficients which are functions of estimates of many elements in the covariance matrix. This reliable estimate of λ_2 is then used in (27) to estimate each component α_i independently of the others.

We have treated two arbitrary means. In conventional communications problems where we deal with equal energy signals which are either anticorrelated or orthogonal, some simplification can result. For anticorrelated signals, for example, the mean is zero, a priori. For orthogonal signals $\boldsymbol{\alpha}$ is perpendicular to $\boldsymbol{\mu}$.

The sample mean and the sample covariance matrix converge for all γ, where $\gamma = |\boldsymbol{\alpha}|/\sigma$. For large γ, as in the univariate case, more rapid convergence can be achieved with nonlinear estimation techniques. We will not delve into this subject now, but might just point out that, for large γ, initial estimation of the partitioning hyperplane based upon the first few samples serves to divide these samples into two distinct relatively tight clumps. Essentially then invoking a supervised-type solution, the mean of each clump is estimated and the perpendicular bisecting plane (of the line of means) is taken as the boundary. Subsequent samples are classified, and then used to up-date the estimates.

V. GENERAL COVARIANCE MATRIX

Having treated the case where the category distributions were spherically symmetric, we now briefly touch upon the ellipsoidally symmetric case, i.e., the component covariance matrices are general. For category distributions which are normal or which are of a form whereby the sample mean and sample covariance matrix converge, the following techniques are applicable. The optimal partitioning hyperplane is defined by the mean $\boldsymbol{\mu}$, and by its perpendicular vector $\mathbf{v}$, where $\mathbf{v} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}$. Therefore, in general, we need to determine these parameters. We examine different cases of varying generality. For all of them we need the sample mean $\hat{\boldsymbol{\mu}}$.

When $\boldsymbol{\Sigma}$ is known and $\mathbf{u}_1$ is known (or is zero), then $\hat{\boldsymbol{\mu}}$ is enough, since $\boldsymbol{\alpha} = (\boldsymbol{\mu} - \mathbf{u}_1)$. For the remaining cases, we need, however, to determine $\hat{\mathbf{B}}$, although not necessarily completely.

When $\mathbf{\Sigma}$ is known, then from (19), $\alpha_i^2 = b_{ii} - \sigma_{ii}$. Choosing α_1 positive, the sign of α_i is obtained from $\alpha_i\alpha_1 = b_{i1} - \sigma_{i1}$. Other combinations of (19) could also be used for solution here.

When only $\mathbf{u}_1$ is known (or is zero), then $\boldsymbol{\alpha}$ is found from $\boldsymbol{\alpha} = \boldsymbol{\mu} - \mathbf{u}_1$, and $\sigma_{ij} = b_{ij} - \alpha_i\alpha_j$.

We now consider that $\mathbf{\Sigma}$, $\mathbf{u}_1$, and $\mathbf{u}_2$ are all unknown. When the component random variates of the component distributions are uncorrelated, i.e., when $\mathbf{\Sigma}$ is diagonal, then $\boldsymbol{\alpha}$ is determined as in (29) and (30), and $\sigma_{ii} = b_{ii} - \alpha_i^2$.

Consider now that $\mathbf{x}$ represents n samples from a stationary process. Then $\mathbf{\Sigma}$ has only n unique elements, and all elements of the principal diagonal are equal, as are elements of each diagonal parallel to it. That is, letting the elements of the first column be $r_0, r_1, \cdots, r_{n-1}$, we have the relation $\sigma_{ij} = r_{|j-i|}$. To assist in the solution we make use of the fourth central moment for the univariate marginal distribution, say the kth one. Denoting this moment with c_{kk}, we present a solution here for the case where the two component distributions are normal. Noting that the marginal kurtosis in the kth coordinate direction is the ratio of c_{kk}, where

$$c_{kk} = 3\sigma_{kk}^2 + 6\sigma_{kk}\,\alpha_k^2 + \alpha_k^4, \tag{31}$$

to the square of b_{kk}, where

$$b_{kk} = \alpha_k^2 + \sigma_{kk}, \tag{32}$$

we obtain the corresponding γ_k in terms of the kurtosis β_k as in (13). Substitution of $\gamma_k^2 = \alpha_k^2/\sigma_{kk}$ in (32) then yields

$$\sigma_{kk} = \frac{b_{kk}}{1 + \gamma_k^2}. \tag{33}$$

For the stationary process, $\sigma_{kk} = \sigma_{ii} = r_0$, and substituting r_0 in (32), $\alpha_i^2 = b_{ii} - r_0$. To simplify notation for the purpose of exposition, assume $k = 1$. Then determine algebraic sign of α_2 and α_3 from examination of

$$\alpha_2(\alpha_1 - \alpha_3) = (b_{21} - b_{32}), \tag{34}$$

obtained from elimination of r_1 in $b_{21} = \alpha_1\alpha_2 + r_1$ and $b_{32} = \alpha_1\alpha_3 + r_1$. Choosing α_1 positive, the signs of α_2 and α_3 become apparent in (34). Then r_1 is determined, with which the signs of the remaining $\{\alpha_i\}$ are determined from $\{b_{ij}\}$, for $i = j + 1$. Now the remaining $\{r_j\}$ can be determined from $\{b_{1j}\}$.

Finally, consider a fully general matrix $\mathbf{\Sigma}$. From (31), (32), and (33), σ_{ii} is obtained, and then from (32) the magnitude of α_i is determined. However, solution of (19) for σ_{ij}, $i \neq j$, and for the algebraic sign of α_i is not unique. Additional relationships must be invoked to obtain unique solution for these latter quantities.

Hill-Climbing

We point out an empirical method for obtaining the optimal plane. Consider all possible hyperplanes passing through the over-all mean. Consider a particular plane defined by its perpendicular vector $\mathbf{v}_d$ having (univariate) kurtosis β_d, corresponding to γ_d. The optimum plane corresponding to vector $\mathbf{v}$ has a signal-to-noise ratio γ which is maximal, and therefore a kurtosis which is minimal. Therefore, using a hill-climbing (or in this case, descending) technique, β_d is evaluated for a particular $\mathbf{v}_d$, and the hyperplane orientation is altered so that β_d is decreased, until the minimum is found. This procedure is not confined to the normal distributions, since we make use of the fact only that β is a monotonically decreasing function of γ, and knowledge of the actual functional relationship is not needed. The hill-climbing technique could of course be much simplified for the spherically symmetric cases, whereby b_{dd} is evaluated and the hyperplane reoriented until the maximum is found.

VI. OTHER MONOTONE DISTRIBUTIONS

As shown elsewhere (Cooper, 1962, 1963), the hyperplane is the boundary form optimally partitioning the sample space when the two multivariate distributions are ellipsoidally symmetric and monotone, and differ from each other only in location. A monotone distribution, of which the normal is one example, is one whose density function is unimodal and monotonically decreasing away from the central location, which for such a symmetric distribution coincides with the mean, median, and mode. So long as the distributions are ellipsoidal monotone, the hyperplane boundary is defined solely in terms of the individual means and the common covariance matrix. The hyperplane passes through the point $\mathbf{u}$, and is defined by the vector $\mathbf{v} = \mathbf{\Sigma}^{-1}(\mathbf{u}_2 - \mathbf{u}_1)$. Much of the discussion of the preceding sections carries forth then to monotone distributions which are spherically or ellipsoidally symmetric, the univariate case of Section III being included.

Actually, the treatment of the last sections applies to general monotone distributions with some qualifications. Included among the classes of

monotone distributions are ones for which the second moment does not exist and even for which the mean exists only in a restricted sense. Such distributions, described in terms of location and scale parameters, nevertheless still lead to a hyperplane boundary. Examples of such distributions are multivariate extensions of the Pearson Type VII distributions of low power parameter, a special case of which is the univariate Cauchy distribution. For these latter distributions, the location and scale parameters can be determined with the methods of order statistics; as, for example, determining the location with the sample median. There are cases, also, where the moments of interest exist but where the corresponding sample moments do not converge, and, here too, recourse is had to estimation methods involving order statistics, as described in the aforementioned references.

Certain of the discussions of the previous sections were developed solely for normal distributions; e.g., the maximum-likelihood development comprising Eqs. (9) and (10) and the first two Appendices. The specific expressions for kurtosis are also specifically for the normal. However, where the functional forms are known, appropriate relations between γ and β can be used, and even where no such knowledge is available, β can be used as an estimate for deciding whether or not γ is large. For any of the monotone distributions β decreases monotonically with γ from a positive value to 1 when γ approaches infinity. The initial value of β for $\gamma = 0$ is 1.8 for a rectangular distribution, 3 for normal, and could be infinite (as, for example, for Pearson Type VII with low power parameter). For a number of types of distributions, such as Pearson Type II, when n is large the marginals are essentially normal, and the earlier kurtosis expression could be used directly. In any event, for any of the distributions, a β of 1.3, for example, would correspond to high γ.

A spherically symmetric distribution with finite variance can represent, for example, a signal $\mathbf{u}_j$ corrupted with additive noise, represented by the spherical distribution with zero mean. The n components of a vector $(\mathbf{x} - \mathbf{u}_j)$ are uncorrelated, and can represent uniformly spaced time samples of the noise process which is stationary and band-limited at twice the reciprocal of the sampling interval. The correlation function of such a noise process is the sinc function, and the power spectral density is band-limited and white. This correlation and spectral property is not, incidentally, confined to monotone noise, and depends only upon the spherical symmetry.

VII. GENERAL DISTRIBUTIONS

In the previous sections, it was shown that rather extensive classes of probabilistic pattern recognition models were amenable to nonsupervised adaptive detection. We now show that there is a general class, which includes the ones already treated, for which similar results are true. This is the class for which the two category prior probabilities are equal and for which the distribution functions of the two categories differ only in location. The partitioning of the sample space appropriate to this class will in most cases involve complicated boundaries. Only in certain cases, including those discussed previously, will the partitioning boundary be a hyperplane.

In the interest of clarity, the proofs and discussion for this section deal only with 1-dimensional sample spaces. However, similar results can be shown to hold in n-dimensional spaces. Denote the cumulative distribution functions for categories 1 and 2 by $F(x + \alpha)$ and $F(x - \alpha)$, respectively. (In this section the term distribution function refers to the cumulative distribution function and is defined for a random variable z by $F(x) = P\{z < x\}$, where $P\{\ \}$ denotes the probability of the event within the brackets.) $F(x)$ is the distribution function for a random variable having mean μ, and 2α is the relative translation of distributions 1 and 2. Since it is not known whether an observation is one of category 1 or category 2, it can be considered as an observation of a population having distribution function $G(x)$, where

$$G(x) = (\tfrac{1}{2})[F(x + \alpha) + F(x - \alpha)].$$

The statistical properties of sequences of independent observations of this population are the properties of the sequence of independent identically distributed random variables $\{z_i\}$, $i = 1, 2, 3, \cdots$, where z_i has the distribution function $G(x)$. We shall be dealing with real-valued functions $F_m(x)$ and $G_m(x)$ and with complex-valued functions $g_m(s)$ of the real variables x and s. These functions are determined by the first m members of $\{z_i\}$.

In terms of $\{z_i\}$, $F(x)$, $G(x)$, and the sequences $\{F_m(x)\}$ and $\{g_m(s)\}$, our first theorem is:

THEOREM 1. *If α is known, the sequence of random variables $\{z_i\}$ uniquely determines a sequence of random functions $\{F_m(x)\}$ having the property that*

$$P\{\lim_{m\to\infty} \sup_{-\infty<x<\infty} |F_m(x) - F(x)| = 0\} = 1.$$

Note that $F(x)$ is completely general here; no restrictions are placed upon it.

Let $f(s)$ be the characteristic function associated with the distribution function $F(x)$. Our second theorem is:

THEOREM 2. *The sequence of random variables $\{z_i\}$ uniquely determines a sequence of random complex functions $\{g_m(s)\}$ having the property that for any bounded interval $[-S, S]$ on the real line*

$$P\{\lim_{m\to\infty} \sup_{s\in[-S,S]} |\cos(s\alpha)f(s) - g_m(s)| = 0\} = 1.$$

The above proofs are constructive, i.e., specific estimators are exhibited. It is helpful to use the engineering feeling developed for stochastic processes and view the sequence $\{z_i\}$ as a discrete parameter stochastic process. Then a sequence of observations is a sample function of the process and is identified, by an element ω of the underlying probability space, as $z_1(\omega)$, $z_2(\omega)$, $\cdots$, $z_m(\omega)$, $\cdots$. Correspondingly, the previously introduced functions $F_m(x)$, $G_m(x)$, and $g_m(s)$ determined by the sample function identified by ω will be denoted $F_m(\omega; x)$, $G_m(\omega; x)$, and $g_m(\omega; s)$. Theorem 1, for example, can be restated as:

If α is known, then almost every sample function $\{z_m(\omega)\}$ uniquely determines a sequence $\{F_m(\omega; x)\}$ which converges to $F(x)$ uniformly in x. A similar restatement of Theorem 2 can be made.

The engineering interest in a result such as Theorem 1 is that it implies the weaker result: If α is known, then for any $\epsilon, \eta > 0$, there exists an integer M such that $m > M$ implies

$$P\{\sup_{-\infty<x<\infty} |F_m(x) - F(x)| < \epsilon\} > 1 - \eta.$$

Hence, if α is known, by taking a sufficiently large number of observations of the over-all population, $F(x)$ can be arbitrarily closely estimated.

We shall show heuristically how in many cases of interest use can be made of Theorem 2 for estimating an unknown α. Corresponding to $F(x + \alpha) + F(x - \alpha) = 2G(x)$ is an equation of characteristic functions, namely, $e^{-i\alpha s}f(s) + e^{i\alpha s}f(s) = 2g(s)$, which reduces to $\cos(\alpha s)f(s) = g(s)$, where $g(s)$ is the characteristic function associated with $G(x)$. In order that α be uniquely determined by $g(s)$, it is necessary that consideration be restricted to a class $\mathfrak{F}$ of distribution functions having the property that if $f(s)$ and $\bar{f}(s)$ are characteristic functions for distribution functions belonging to $\mathfrak{F}$ then

$$\cos(\alpha s)f(s) = \cos(\bar{\alpha}s)\bar{f}(s)$$

implies $\bar{\alpha} = \alpha$. Equivalently, it is necessary that consideration be restricted to a class $\mathfrak{F}$ having the property that if $F(x)$ and $\bar{F}(x)$ belong to $\mathfrak{F}$, then

$$\bar{F}(x + \bar{\alpha}) + \bar{F}(x - \bar{\alpha}) = F(x + \alpha) + F(x - \alpha) \text{ implies } \bar{\alpha} = \alpha.$$

It is apparent that $g(s)$ has zeroes wherever $\cos(\alpha s)$ has zeroes. If the zeroes of $f(s)$ are neither periodic nor too many, then examination of the zeroes of $g(s)$ can provide a simple method for the determination of α. As $g(s)$ is not known a priori, Theorem 2 is brought to bear, and in many cases of interest examination of $g_m(s)$ will provide an estimator which converges to α. If $F(x)$ is continuous, then an estimator convergent to α with probability 1 can be used in place of α in Theorem 1.

Before proceeding with the proofs, an application of the preceding theorems to optimum detector or sample space partition estimation is briefly pointed out. Suppose α were known, $F(x)$ were continuous, and, as would be the case in most problems of engineering interest, the optimum partition for the line consisted of a finite number, say K or fewer, of points. Then the partitioning points occur at the extrema of the function $F(x - \alpha) - F(x + \alpha)$. An estimator $F_m(x - \alpha) - F_m(x + \alpha)$ uniformly convergent in x is constructible by Theorem 1. Examination of the points of occurrence of the extrema of this function provides a convergent K point estimator for the optimum partition.

We proceed with a proof of Theorem 1. Let $G_m(\omega; x)$ be the empirical distribution function of the m statistically independent observations $z_1(\omega), \cdots, z_m(\omega)$ of the population having distribution function $G(x)$. By this is meant that

$$G_m(\omega, x) = \frac{1}{m} \sum_{i=1}^{m} D[x - z_i(\omega)]$$

$$\text{where} \quad D[x - z_i(\omega)] = 0 \quad \text{if} \quad x \leqq z_i(\omega)$$

$$= 1 \quad \text{if} \quad x > z_i(\omega).$$

Thus, $G_m(\omega; x)$ is a step function having positive jumps of size $1/m$ at the observations. By the Glivenko-Cantelli Theorem,

$$P\{\lim_{m \to \infty} \sup_{-\infty < x < \infty} |G_m(x) - G(x)| = 0\} = 1; \tag{35}$$

$G_m(x)$ is a convergent estimator for $G(x)$. The structure of the remainder of the proof consists of first showing that if $G(x)$ is generated by some

unknown distribution function $F(x)$, then, given $G(x)$, a solution of the equation

$$\bar{F}(x + \alpha) + \bar{F}(x - \alpha) = 2G(x) \tag{36}$$

for $F(x)$ can be found. A function $H_m(\omega; x)$ is obtained by replacing all occurrences of $G(x)$ in the solution of (36) by $G_m(\omega; x)$, and it is then shown that the convergence of $G_m(\omega; x)$ to $G(x)$ (Eq. (35)) as m becomes large ensures the convergence of $H_m(\omega; x)$ to $F(x)$. Finally, $H_m(\omega; x)$ is modified and a convergent estimate $F_m(\omega; x)$, which is itself a distribution function and satisfies Theorem 1, is arrived at.

LEMMA 1. *If α is known, then*

$$\sum_{k=1}^{\infty}(-1)^{k+1}2G[x - (2k - 1)\alpha] \tag{37}$$

converges to the distribution function $F(x)$ which gave rise to $G(x)$. The convergence is uniform in x on left semi-infinite intervals (i.e. intervals of the form $(-\infty, \xi]$ with ξ finite).

PROOF: Choose any point ξ on the real line. Since $F(x)$ is a solution of Eq. (36), it follows that

$$\begin{aligned}\sum_{k=1}^{j}(-1)^{k+1}2G[x - (2k - 1)\alpha] &= \sum_{k=1}^{j}(-1)^{k+1}\{F[x + \alpha - (2k - 1)\alpha] + F[x - \alpha - (2k - 1)\alpha]\} \\ &= F(x) + (-1)^{j+1}F(x - 2j\alpha).\end{aligned}$$

But $F(x)$ a distribution function implies that $F(x - 2j\alpha) \to 0$ as $j \to \infty$. Hence, (37) converges to $F(x)$ uniformly in x, $x \in (-\infty, \xi]$.

A similar convergent series of translates of $G_m(\omega; x)$ provides an estimator for $F(x)$, namely,

$$H_m(\omega; x) = \sum_{k=1}^{\infty}(-1)^{k+1}2G_m[\omega; x - (2k - 1)\alpha]. \tag{38}$$

Function (38) is well defined since for each x only a finite number of summands of (38) are nonzero. Other characteristics pertinent to the proof are that (38) is bounded above by its first term, $2G_m[\omega; x - \alpha]$, and is nonnegative since $G_m(\omega; x)$ is a distribution function.

Appendix III contains a proof of the fact that for almost every ω $H_m(\omega; x)$ converges to $F(x)$ uniformly in x $(-\infty < x < \infty)$ as m

becomes large. Let $z^m(\omega)$ denote the largest of the first m observations. An estimator which is a distribution function and satisfies Theorem I is obtained from $H_m(\omega; x)$, as

$$F_m(\omega; x) = \begin{Bmatrix} \min\{\sup_{y \leqq x} H_m(\omega; y), 1\}, & \text{for } x \leqq z^m(\omega); \\ 1, & \text{for } x > z^m(\omega) \end{Bmatrix} \tag{39}$$

Note that for each m and for all x ($-\infty < x < \infty$) only a finite number of summands of (38) appear in (39).

The proof of Theorem 2 is almost immediate.

PROOF OF THEOREM 2. Let

$$g_m(\omega; s) = \int e^{isx}\, dG_m(\omega; x).$$

From the proof of Theorem 1, for almost all ω, $G_m(\omega; x)$ converges to $G(x)$ completely (pointwise, and the variation of $G_m(\omega; x)$ converges to the variation of $G(x)$). Hence (see Loève, 1960, p. 191, Theorems B and C) for such ω, $g_m(\omega; s)$ converges to $g(s)$ uniformly in s on every finite interval $[-S, S]$. That is, for any positive number S,

$$P\{\lim_{m\to\infty} \sup_{s\in[-S,S]} |g_m(s) - g(s)| = 0\} = 1.$$

VIII. CHARACTERISTIC FUNCTIONS

As was illustrated in Section VII, the characteristic function is of value in determining whether knowledge of the over-all distribution is sufficient for uniquely specifying the individual distributions and the optimum decision rule. In addition, the empirical characteristic function, $g_m(\mathbf{s})$, is of use in many situations for estimating some of the parameters necessary for determining the optimum decision rule.

We briefly point out a few of the pertinent relationships involving characteristic functions. Denoting the over-all characteristic function by $g(\mathbf{s})$, the two category characteristic functions by $f_1(\mathbf{s})$ and $f_2(\mathbf{s})$, and the prior probabilities of categories 1 and 2 by q and $(1 - q)$, we have the equation $g(\mathbf{s}) = qf_1(\mathbf{s}) + (1 - q)f_2(\mathbf{s})$. If $p_1(\mathbf{x}) = p_3(\mathbf{x} + \boldsymbol{\alpha})$ and $p_2(\mathbf{x}) = p_3(\mathbf{x} - \boldsymbol{\alpha})$, this becomes $g(\mathbf{s}) = [qe^{-i\boldsymbol{\alpha}'\mathbf{s}} + (1 - q)e^{i\boldsymbol{\alpha}'\mathbf{s}}]f(\mathbf{s})$, where $f(\mathbf{s})$ is the characteristic function associated with $p_3(\mathbf{x})$. Lastly, if $q = \frac{1}{2}$, the above equation reduces to $g(\mathbf{s}) = \cos(\boldsymbol{\alpha}'\mathbf{s})f(\mathbf{s})$.

IX. UNEQUAL PRIOR PROBABILITIES

The linear statistical estimation procedures can be simply modified to treat the case for which the prior probabilities are unequal, where the

a priori probability of category 1 is q. We illustrate this here for the spherical distributions of Section IV. The decision boundary parameters are defined in terms of the four distribution parameters $\mathbf{u}_1$, $\mathbf{u}_2$, σ, and q, which can be readily estimated in terms of central sample moments. By any of the previous methods, we estimate the over-all mean $\boldsymbol{\mu}$ and the two eigenvalues and the principal eigenvector of the over-all covariance matrix $\mathbf{B}$. The normalized principal eigenvector is denoted with $\mathbf{e}_p$, where $|\mathbf{e}_p| = 1$. Denoting components in the direction of $\mathbf{e}_p$ with subscript p, the component of a sample $\mathbf{x}$ in this direction is $y_p = \mathbf{x}'\mathbf{e}_p$. Representing the third central moment with ϕ, we make use of the following four equations:

$$\lambda_2 = \sigma^2 \tag{40}$$

$$\mu_p = qu_{1p} + (1 - q)u_{2p} \tag{41}$$

$$\lambda_1 = \sigma^2 + q(1 - q)(u_{2p} - u_{1p})^2 \tag{42}$$

$$\phi_p = q(1 - q)(1 - 2q)(u_{2p} - u_{1p})^3. \tag{43}$$

In terms of these the parameters of interest are:

$$\sigma = \sqrt{\lambda_2} \tag{44}$$

$$q = \frac{1}{2}\left\{1 \pm \left[1 - \left[1 + \left(\frac{\phi_p^2}{4(\lambda_1 - \lambda_2)^3}\right)\right]^{-1}\right]^{1/2}\right\} \tag{45}$$

$$u_{1p} = \mu_p - (1 - q)[(\lambda_1 - \lambda_2)/q(1 - q)]^{1/2} \tag{46}$$

$$u_{2p} = \mu_p + q[(\lambda_1 - \lambda_2)/q(1 - q)]^{1/2}. \tag{47}$$

In practice we of course use estimates, i.e., $\hat{\lambda}_1$, $\hat{\lambda}_2$, $\hat{\mu}_p$, and $\hat{\phi}_p$.

For normal distributions the minimal-error-probability partition is a hyperplane defined by its perpendicular vector $\mathbf{e}_p$ and including the point $\mathbf{s} = \boldsymbol{\mu} - \theta\mathbf{e}_p$, where

$$\theta = \frac{\sigma^2}{(u_{2p} - u_{1p})} \log\left(\frac{1 - q}{q}\right). \tag{48}$$

An alternative and simpler solution, along the lines of Eqs. (29) and (30), is obtained directly from some of the elements of $\mathbf{B}$ and from ϕ_1, the univariate third central moment in the first coordinate direction. Taking $\mathbf{t} = \mathbf{u}_2 - \boldsymbol{\mu}$ and $\mathbf{s} = \mathbf{u}_1 - \boldsymbol{\mu}$, and defining $K_1 = b_{31}b_{21}/b_{32}$, we obtain: $t_1 = (\phi_1/2K_1) \pm [(\phi_1/2K_1)^2 + K_1]^{1/2}$; $t_j/t_1 = b_{jk}/b_{1k}$, $k \neq j$ or 1; $q = t_1^2/(K_1 + t_1^2)$; $\sigma^2 = b_{11} - K_1$; $\mathbf{s} = -\mathbf{t}(1 - q)/q$, this latter relation corresponding to $\mathbf{t}$ and $\mathbf{s}$ being the $+$ and $-$ solutions to the first equation.

X. MULTIPLE-CATEGORY

Briefly examining the multiple-category case, suppose we have J spherically symmetric (normal) distributions differing only in location. Equations can be set up in terms of the estimated parameters for obtaining the hyperplane partitions. Actually the information to be determined is the location of the means themselves. Suppose the J means define a $(J - 1)$-dimensional hyperplane, i.e., the difference vectors between one of the means with each of the others are linearly independent. We begin by obtaining the sample mean and the sample covariance matrix of the over-all "multimodal" distribution. All of the discriminatory information is contained in the hyperplane defined by the means. The $n \times n$ covariance matrix will have positive real eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_J$, listed in decreasing value, where the smallest one, λ_J, is of multiplicity $(n - J + 1)$. The hyperplane of the means is defined by the eigenvectors corresponding to the $(J - 1)$ largest eigenvalues.

Suppose we did not know J. It can be determined by finding the smallest eigenvalue and evaluating its multiplicity. There are a number of well known techniques for finding λ_J and its multiplicity, e.g., an iterative method on the inverse of the matrix leads to λ_J. Repeated iterations or testing of derivatives of the characteristic polynomial will reveal the multiplicity. Actually, by this method, we can determine the number of categories present, from a large sample set, when the component distributions differ among themselves only in location and are all spherically symmetric (not necessarily monotone), and have finite fourth moments. We postpone further treatment of the nonsupervised multiple-category case to a future paper.

XI. TIME-VARYING STATISTICS

There are situations where the statistics of the categories change with time. It is then desirable to have an adaptive detector which will keep pace with these changes and remain near optimum. In the following paragraph we illustrate, with an example, a procedure for converting previously discussed nonsupervised adaptive detectors to ones suitable for slowly time-varying statistics. While the statistics are time-varying slowly the detector will track and remain near optimum, and if sudden changes occur the detector will commence adaptation immediately.

Tracking capability can be achieved in several ways which we illustrate in terms of modification of the sample mean. For example, Eq. (6) can

be modified so that the summation is taken over the K most recent samples. A drawback of this procedure is that K samples must at all times be stored. To circumvent this, an exponential type of weighting can be used. Two modifications of (8) which achieve this are: replace m with $m\theta$, where $0 < \theta < 1$, to obtain $\hat{\mu}_{m+1} = (m\theta\hat{\mu}_m + z_{m+1})/(m\theta + 1)$; or replace m with K (made only if $m \geqq K$) to obtain $\hat{\mu}_{m+1} = (K\hat{\mu}_m + z_{m+1})/(K + 1)$. The last two procedures allow for continuous updating of the estimate, yet need only the limited memory requirement of the usual sample mean. Replacement of m with K is good for the time-varying case, where K can be chosen to allow for satisfactory tracking. The $m\theta$ substitution effects a weighting which is significant for a progressively greater number of samples, and therefore, although not well suited for the time-varying case, is good for the stationary case where we wish to eliminate the initial transients in the "learning," as discussed at the end of Section III.

XII. REMARKS

Some of the techniques which have been presented are simple and yet would perform near optimally. Others are presented as existence statements to show that there is adequate information present to obtain a solution, and a solution is presented, although it may not necessarily be especially simple to implement. Space limitations prevented a more exhaustive treatment of some of the topics, such as the case of different prior probabilities for Section VII. Various ramifications of some of the ideas discussed could undoubtedly lead to improvements.

For the cases treated where the hyperplane boundary is optimum for the minimum-error-probability criterion, it also optimally satisfies the Minimax criterion, because of the symmetries involved.

In classic signal detection studies, one usually demonstrates the virtue of the correlation detector by showing that whereas the input signal-to-noise ratio is small, the output signal-to-noise ratio is large, where the usually defined signal-to-noise ratio is equivalent to our γ^2. $(S/N)_{\text{in}}$ is then $(\boldsymbol{\alpha}'\boldsymbol{\alpha}/n)/\sigma^2$, and $(S/N)_{\text{out}}$ is $\gamma^2 = \boldsymbol{\alpha}'\boldsymbol{\alpha}/\sigma^2$. The convergence of some nonlinear nonsupervised adaptive procedures might be conditioned upon γ^2 being large. The techniques we have presented cover all values for γ. For example, the sample mean converges regardless how small γ is (in fact better for the small values), even when $(S/N)_{\text{in}}$ is so small that γ is perhaps not large enough to give adequate detection reliability. Nevertheless, for such a miniscule γ case, the (nonsupervised) sample

mean still converges to the optimum threshold. With suitable redundant coding, even the miniscule γ case can allow high detector reliability.

Various NASD (Nonsupervised Adaptive Signal Detector) Systems are described in (D. Cooper and P. Cooper, 1964), as well as is further discussion on the philosophical motivation. The systems described there make use of the nonlinear procedures and the iterative technique referred to in Section IV for determination of the principal eigenvector.

The authors have each continued different aspects of this research, and in particular some of David Cooper's extensions of this work comprise part of his doctoral dissertation.

ACKNOWLEDGMENT

David Cooper takes pleasure in acknowledging valuable discussions he has had on this subject with his thesis adviser Professor Ralph J. Schwarz.

APPENDIX I. MAXIMUM-LIKELIHOOD ESTIMATORS

The maximum-likelihood estimators for the parameters of the univariate bimodal distribution of (5) are determined from the multiple log likelihood function

$$\psi = \sum_{k=1}^{m} \log L(z_k) = -m \log(\sigma\sqrt{2\pi}) - m\alpha^2/2\sigma^2 - (1/2\sigma^2)\sum_{k=1}^{m}(z_k - \mu)^2 + \sum_{k=1}^{m} \log(\cosh[(\alpha/\sigma^2)(z_k - \mu)]). \tag{49}$$

Taking the partial derivatives of ψ with respect to μ, α, and σ, and setting them equal to zero leads, respectively, to the following three equations:

$$\hat{\mu} = \frac{1}{m}\sum_{k=1}^{m} z_k - \frac{\alpha}{m}\sum_{k=1}^{m} \tanh\,[(\alpha/\sigma^2)(z_k - \hat{\mu})], \tag{50}$$

$$\hat{\alpha} = \frac{1}{m}\sum_{k=1}^{m} (z_k - \mu)\,\tanh\,[(\hat{\alpha}/\sigma^2)(z_k - \mu)], \tag{51}$$

$$\hat{\sigma}^2 = \frac{1}{m}\sum_{k=1}^{m} (z_k - \mu)^2 + \alpha^2 - \frac{2\alpha}{m}\sum_{k=1}^{m} (z_k - \mu)\,\tanh\,[(\alpha/\hat{\sigma}^2)(z_k - \mu)]. \tag{52}$$

If any two parameters are known, the maximum-likelihood estimator for the third is obtained from solution of the single equation appropriate to that estimator, wherein the estimated parameter is indicated with a

circumflex. If all three parameters are unknown then these parameters are all overscored with circumflex in the three equations, simultaneous solution of which gives the maximum-likelihood estimators for the three parameters. For the latter case (only), immediate substitution reduces (52) to

$$\hat{\sigma}^2 = \frac{1}{m}\sum_{k=1}^{m}(z_k - \hat{\mu})^2 - \hat{\alpha}^2. \tag{53}$$

APPENDIX II. ASYMPTOTIC VARIANCE

As is well known, the asymptotic variance of the maximum-likelihood estimator for a parameter is the negative reciprocal of the expectation of the second derivative of ψ with respect to that parameter.

$$\frac{\partial\psi}{\partial\mu} = \frac{1}{\sigma^2}\left\{\sum_{k=1}^{m}(z_k - \mu) - \alpha\sum_{k=1}^{m}\tanh\left[(\alpha/\sigma^2)(z_k - \mu)\right]\right\}, \tag{54}$$

$$\frac{\partial^2\psi}{\partial\mu^2} = -\frac{m}{\sigma^2}\left\{1 - (\alpha/\sigma)^2\frac{1}{m}\sum_{k=1}^{m}\operatorname{sech}^2\left[(\alpha/\sigma^2)(z_k - \mu)\right]\right\}. \tag{55}$$

Then the asymptotic variance of the maximum-likelihood estimator of μ (under the assumption that α and σ are known) is

$$\text{Var} = -\left[E\left(\frac{\partial^2\psi}{\partial\mu^2}\right)\right]^{-1} = \frac{\sigma^2}{m}\left[1 - Q\gamma^2\right]^{-1} \tag{56}$$

where $\gamma = |\alpha|/\sigma$, and

$$\begin{aligned} Q &= E(\operatorname{sech}^2\left[(\alpha/\sigma^2)(x - \mu)\right]) \\ &= \int_{-\infty}^{\infty} p(x)\operatorname{sech}^2\left[(\alpha/\sigma^2)(x - \mu)\right]dx \\ &= e^{-\gamma^2/2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\gamma}e^{-y^2/2\gamma^2}\operatorname{sech}(y)\,dy, \end{aligned} \tag{57}$$

and where $p(x)$ is given in (5). It is a simple matter to obtain directly an upper and a lower bound for Q, as a function of γ, wherein the two bounds are related in that one is twice the other. However, computer solution for Q has been substituted into (56) to obtain the curve shown in Fig. 1.

APPENDIX III. PROOF OF THEOREM 1

LEMMA 2. *If α is known, then $H_m(x)$ converges to $F(x)$ uniformly in x on left semi-infinite intervals with probability* 1.

PROOF: Choose any $\epsilon > 0$ and any point ξ on the real line. Choose an integer M such that

$$2G[\xi - (2M - 1)\alpha] < \epsilon/4. \tag{58}$$

Equation (35) implies that for almost every ω taking m sufficiently large ensures that

$$[\sup_{-\infty<x<\infty} 2 \mid G(x) - G_m(\omega; x) \mid] < \epsilon/4M. \tag{59}$$

Now

$$\mid F(x) - H_m(\omega; x) \mid$$

$$\leqq \left| F(x) - \sum_{k=1}^{\infty} (-1)^{k+1} 2G[x - (2k - 1)\alpha] \right| \tag{60}$$

$$+ \left| \sum_{k=1}^{\infty} (-1)^{k+1} 2G[x - (2k - 1)\alpha] - \sum_{k=1}^{\infty} (-1)^{k+1} 2G_m[\omega; x - (2k - 1)\alpha] \right|. \tag{61}$$

Function (60) converges to 0 uniformly on $(-\infty, \xi]$ by Lemma 1, and (61) is bounded above by

$$\left| \sum_{k=1}^{M-1} (-1)^{k+1} 2\{G[x - (2k - 1)\alpha] - G_m[\omega; x - (2k - 1)\alpha]\} \right| \tag{62}$$

$$+ \left| \sum_{k=M}^{\infty} (-1)^{k+1} 2G[x - (2k - 1)\alpha] - \sum_{k=M}^{\infty} (-1)^{k+1} 2G_m[\omega; x - (2k - 1)\alpha] \right|. \tag{63}$$

Inequality (59) implies that (62) is less than $\epsilon/4$. The magnitudes of the first and second summations of (63) are bounded above by $2G[x - (2M - 1)\alpha]$ and $2G_m[\omega; x - (2M - 1)\alpha]$, respectively. But if x belongs to $(-\infty, \xi]$, then as a result of (58) and (59) it follows that

$$2G[x - (2M - 1)\alpha] < \epsilon/4$$

and

$$2G_m[\omega; x - (2M - 1)\alpha] < \epsilon/4M + \epsilon/4 \leqq \epsilon/2.$$

Hence, (63) is less than $3\epsilon/4$, and (61) is therefore bounded above by ϵ. Since this is true for almost every ω, the Lemma is proved.

The extension of this result to uniform convergence in x over the interval $(-\infty, \infty)$ follows simply. For, let $\epsilon > 0$ be arbitrary. Consider any ω for which $G_m(\omega; x) \to G(x)$, $(-\infty < x < \infty)$, as $m \to \infty$. Choose a point ξ for which

$$1 - G(\xi) < \epsilon/6. \tag{64}$$

Lemma 2 and the convergence of $G_m(\omega; x)$ to $G(x)$ permit the choice of an integer M such that $m \geqq M$ implies

$$|G(x) - G_m(\omega; x)| < \epsilon/6 \text{ for all } x, \tag{65}$$

and

$$|H_m(\omega; y) - F(y)| < \epsilon/3 \text{ for } y \in (-\infty, \xi + 4\alpha]. \tag{66}$$

Let $m \geqq M$. Now

$$\begin{aligned} H_m(\omega; x) = \sum_{k=1}^{K} (-1)^{k+1} 2G_m[\omega; x - (2k-1)\alpha] \\ + \sum_{k=K+1}^{\infty} (-1)^{k+1} 2G_m[\omega; x - (2k-1)\alpha] \end{aligned} \tag{67}$$

with K an even integer satisfying

$$\xi < x - (2K-1)\alpha < \xi + 4\alpha. \tag{68}$$

The first summation of (67) is positive and is bounded above by $2G_m[\omega; x - \alpha] - 2G_m[\omega; x - (2K-1)\alpha]$. In view of (64), (65), and (68), this is itself bounded above by $2\epsilon/3$. Making the change of variable $y = x - 2K\alpha$, we see that the second summation of Eq. (67) is $H_m(\omega; y)$, with $y < \xi + 4\alpha$. Hence, the second summation satisfies (66) and thus leads to

$$|H_m(\omega; x) - F(x)| < 2\epsilon/3 + \epsilon/3 = \epsilon.$$

This is true for any x, as was to be proved.

RECEIVED: November 4, 1963.

REFERENCES

COOPER, P. W. (1962), The hyperplane in pattern recognition. *Cybernetica* **5**, 215–238.

COOPER, P. W. (1963), Hyperplanes, hyperspheres, and hyperquadrics as decision boundaries. *ONR-COINS Symp.* at Northwestern University (June, 1963), appearing in "Computers and Information Sciences." Spartan Books, Washington.

COOPER, D. B. AND COOPER, P. W. (1964), Adaptive pattern recognition and signal detection without supervision. *IEEE Intern. Conv. Record,* Pt. I.

FADDEEVA, V. N. (1959), "Computational Methods of Linear Algebra." Dover, New York.

HINICH, M. J. (1962), A model for a self-adapting filter. *Inform. Control* **5,** 185–203.

JAKOWATZ, C. V., SHUEY, R. L., AND WHITE, G. M. (1961), Adaptive waveform recognition. *In* "Information Theory," C. Cherry, ed. Butterworths, Washington.

LOÈVE, M. (1960), "Probability Theory," 2nd ed. Van Nostrand, Princeton, New Jersey.

PRICE, R. AND GREEN, P. E. (1958), A communication technique for multipath channels. *PROC. IRE* **46,** 555–570.

SHANNON, C. (1949), Communications in the presence of noise. *PROC. IRE* **37,** 10–21.

WILKS, S. S. (1962), "Mathematical Statistics." Wiley, New York.